Webpage Classification with ACO-enhanced Fuzzy-Rough Feature Selection.

Jensen, Richard; Shen, Qiang

Publication date:
2006

Citation for published version (APA):
Jensen, R., & Shen, Q. (2006). *Webpage Classification with ACO-enhanced Fuzzy-Rough Feature Selection.*. 147-156. http://hdl.handle.net/2160/442

Webpage Classification with ACO-Enhanced Fuzzy-Rough Feature Selection

Richard Jensen and Qiang Shen

Department of Computer Science, The University of Wales, Aberystwyth
{rkj, qqs}@aber.ac.uk

Abstract. Due to the explosive growth of electronically stored information, automatic methods must be developed to aid users in maintaining and using this abundance of information effectively. In particular, the sheer volume of redundancy present must be dealt with, leaving only the information-rich data to be processed. This paper presents an approach, based on an integrated use of fuzzy-rough sets and Ant Colony Optimization (ACO), to greatly reduce this data redundancy. The work is applied to the problem of webpage categorization, considerably reducing dimensionality with minimal loss of information.

1 Introduction

The World Wide Web (WWW) is an information resource, whose full potential may not be realised unless its content is adequately organised and described. However, due to the immense size and dynamicity of the web, manual categorization is not a practical solution to this problem. There is a clear need for automated classification of web content.

Many classification problems involve high dimensional descriptions of input features. It is therefore not surprising that much research has been done on dimensionality reduction [4]. However, existing work tends to destroy the underlying semantics of the features after reduction (e.g. transformation-based approaches) or require additional information about the given data set for thresholding (e.g. entropy-based approaches). A technique that can reduce dimensionality using information contained within the data set and preserving the meaning of the features is clearly desirable. Rough set theory (RST) can be used as such a tool to discover data dependencies and reduce the number of features contained in a dataset by purely structural methods [9]. Given a dataset with discretized attribute values, it is possible to find a subset (termed a *reduct*) of the original attributes using RST that are the most informative; all other attributes can be removed from the dataset with minimal information loss.

Although this is useful, it is more often the case that data is *real-valued*, and this is where traditional rough set theory encounters a problem. In the theory, it is not possible to say whether two attribute values are similar and to what extent they are the same; for example, two close values may only differ as a result of noise, but in RST they are considered to be as different as two values of a different order of magnitude. It is, therefore, desirable to develop these techniques to provide the means

S. Greco et al. (Eds.): RSCTC 2006, LNAI 4259, pp. 147–156, 2006.

of data reduction for crisp and real-value attributed datasets which utilises the extent to which values are similar. This can be achieved through the use of *fuzzy-rough* sets. Fuzzy-rough sets encapsulate the related but distinct concepts of vagueness (for fuzzy sets [17]) and indiscernibility (for rough sets [9]), both of which occur as a result of imprecision, incompleteness and/or uncertainty in knowledge [5].

Ant Colony Optimization (ACO) techniques are based on the behaviour of real ant colonies used to solve discrete optimization problems [1]. These have been successfully applied to a large number of difficult combinatorial problems such as the quadratic assignment and the traveling salesman problems. This method is particularly attractive for feature selection as there seems to be no heuristic that can guide search to the optimal minimal subset (of features) every time. Additionally, it can be the case that ants discover the best feature combinations as they proceed throughout the search space. This paper investigates how ant colony optimization may be applied to the difficult problem of finding optimal feature subsets, using fuzzy-rough sets, for the classification of web content.

The rest of this paper is structured as follows. The second section describes the theory of fuzzy-rough set feature selection. Section 3 introduces the main concepts in ACO and details how this may be applied to the problem of feature selection in general, and fuzzy-rough feature selection in particular. The fourth section describes the system components and experimentation carried out for the purposes of web content classification. Section 5 concludes the paper, and proposes further work in this area.

2 Fuzzy-Rough Feature Selection

The reliance on discrete data for the successful operation of rough set-based feature selection methods such as [2,6,16] can be seen as a significant drawback of the approach. Indeed, this requirement implies an objectivity in the data that is simply not present. For example, in a medical dataset, values such as *Yes* or *No* cannot be considered objective for a *Headache* attribute as it may not be straightforward to decide whether a person has a headache or not to a high degree of accuracy. Again, consider an attribute *Blood Pressure*. In the real world, this is a real-valued measurement but for the purposes of rough set theory must be discretised into a small set of labels such as *Normal*, *High*, etc. Subjective judgments are required for establishing boundaries for objective measurements.

A better way of handling this problem is the use of fuzzy-rough sets [8]. Subjective judgments are not entirely removed as fuzzy set membership functions still need to be defined. However, the method offers a high degree of flexibility when dealing with real-valued data, enabling the vagueness and imprecision present to be modelled effectively. By employing fuzzy-rough sets, it is possible to use this information to better guide feature selection.

2.1 Fuzzy Equivalence Classes

In the same way that crisp equivalence classes are central to rough sets, *fuzzy* equivalence classes are central to the fuzzy-rough set approach [5]. For typical

applications, this means that the decision values and the conditional values may all be fuzzy. The family of normal fuzzy sets produced by a fuzzy partitioning of the universe of discourse can play the role of fuzzy equivalence classes [5].

2.2 Fuzzy Lower and Upper Approximations

The fuzzy lower and upper approximations are fuzzy extensions of their crisp counterparts. Informally, in crisp rough set theory, the lower approximation of a set contains those objects that belong to it with certainty. The upper approximation of a set contains the objects that possibly belong. The definitions given in [5] diverge a little from the crisp upper and lower approximations, as the memberships of individual objects to the approximations are not explicitly available. As a result of this, the fuzzy lower and upper approximations are redefined as:

$$\mu_{\underline{P}X}(x) = \sup_{F \in \mathbb{U}/P} min(\mu_F(x), \inf_{y \in \mathbb{U}} max\{1 - \mu_F(y), \mu_X(y)\}) \tag{1}$$

$$\mu_{\overline{P}X}(x) = \sup_{F \in \mathbb{U}/P} min(\mu_F(x), \sup_{y \in \mathbb{U}} min\{\mu_F(y), \mu_X(y)\}) \tag{2}$$

The tuple $< \underline{P}X, \overline{P}X >$ is called a fuzzy-rough set.

For an individual feature, a, the partition of the universe by $\{a\}$ (denoted $\mathbb{U}/IND(\{a\})$) is considered to be the set of those fuzzy equivalence classes for that feature. For subsets of feature, the following is used:

$$\mathbb{U}/P = \otimes\{a \in P : \mathbb{U}/IND(\{a\})\} \tag{3}$$

Each set in $\mathbb{U}/P$ denotes an equivalence class. The extent to which an object belongs to such an equivalence class is therefore calculated by using the conjunction of constituent fuzzy equivalence classes, say F_i, $i = 1, 2, ..., n$:

$$\mu_{F_1 \cap ... \cap F_n}(x) = min(\mu_{F_1}(x), \mu_{F_2}(x), ..., \mu_{F_n}(x)) \tag{4}$$

2.3 Fuzzy-Rough Reduction Process

Fuzzy-Rough Feature Selection (FRFS) [7] builds on the notion of the fuzzy lower approximation to enable reduction of datasets containing real-valued features. The process becomes identical to the crisp approach when dealing with nominal well-defined features.

The crisp positive region in the standard RST is defined as the union of the lower approximations. By the extension principle, the membership of an object $x \in \mathbb{U}$, belonging to the fuzzy positive region can be defined by

$$\mu_{POS_P(Q)}(x) = \sup_{X \in \mathbb{U}/Q} \mu_{\underline{P}X}(x) \tag{5}$$

Using the definition of the fuzzy positive region, a new dependency function between a set of features Q and another set P can be defined as follows:

$$\gamma'_P(Q) = \frac{|\mu_{POS_P(Q)}(x)|}{|\mathbb{U}|} = \frac{\sum_{x \in \mathbb{U}} \mu_{POS_P(Q)}(x)}{|\mathbb{U}|} \tag{6}$$

As with crisp rough sets, the dependency of Q on P is the proportion of objects that are discernible out of the entire dataset. In the present approach, this corresponds to determining the fuzzy cardinality of $\mu_{POS_P(Q)}(x)$ divided by the total number of objects in the universe.

A new QuickReduct algorithm, based on the crisp version [2], has been developed [7]. It employs the new dependency function γ' to choose which features to add to the current reduct candidate. The algorithm terminates when the addition of any remaining feature does not increase the dependency.

Conventional hill-climbing approaches to feature selection often fail to find maximal data reductions or minimal reducts. Some guiding heuristics are better than others for this, but as no perfect heuristic exists there can be no guarantee of optimality. When maximal data reductions are required, other search mechanisms must be employed. Although these methods also cannot ensure optimality, they provide a means by which the best feature subsets might be found. This motivates the development of feature selection based on ant colony optimization.

3 Ant Colony Optimization-Based Feature Selection

3.1 Swarm Intelligence

Swarm Intelligence (SI) is the property of a system whereby the collective behaviours of simple agents interacting locally with their environment cause coherent functional global patterns to emerge [1]. SI provides a basis with which it is possible to explore collective (or distributed) problem solving without centralized control or the provision of a global model. For example, ants are capable of finding the shortest route between a food source and their nest without the use of visual information and hence possess no global world model, adapting to changes in the environment. Those SI techniques based on the behaviour of ant colonies used to solve discrete optimization problems are classed as Ant Colony Optimization (ACO) techniques [1].

The ability of real ants to find shortest routes is mainly due to their depositing of pheromone as they travel; each ant probabilistically prefers to follow a direction rich in this chemical. The pheromone decays over time, resulting in much less pheromone on less popular paths. Given that over time the shortest route will have the higher rate of ant traversal, this path will be reinforced and the others diminished until all ants follow the same, shortest path (the "system" has converged to a single solution). It is also possible that there are many equally short paths.

ACO is particularly attractive for feature selection as there seems to be no heuristic that can guide search to the optimal minimal subset every time. Additionally, it can be the case that ants discover the best feature combinations as they proceed throughout the search space.

3.2 Feature Selection

The feature selection task may be reformulated into an ACO-suitable problem. ACO requires a problem to be represented as a graph - here nodes represent

features, with the edges between them denoting the choice of the next feature. The search for the optimal feature subset is then an ant traversal through the graph where a minimum number of nodes are visited that satisfies the traversal stopping criterion.

A suitable heuristic desirability of traversing between features could be any subset evaluation function - for example, an entropy-based measure [10] or the fuzzy-rough set dependency measure. Depending on how optimality is defined for the particular application, the pheromone may be updated accordingly. For instance, subset minimality and "goodness" are two key factors so the pheromone update should be proportional to "goodness" and inversely proportional to size. How "goodness" is determined will also depend on the application. In some cases, this may be a heuristic evaluation of the subset, in others it may be based on the resulting classification accuracy of a classifier produced using the subset.

The heuristic desirability and pheromone factors are combined to form the so-called probabilistic transition rule, denoting the probability of an ant k at feature i choosing to move to feature j at time t:

$$p_{ij}^k(t) = \frac{[\tau_{ij}(t)]^\alpha.[\eta_{ij}]^\beta}{\sum_{l \in J_i^k}[\tau_{il}(t)]^\alpha.[\eta_{il}]^\beta} \tag{7}$$

where J_i^k is the set of ant k's unvisited features, η_{ij} is the heuristic desirability of choosing feature j when at feature i and $\tau_{ij}(t)$ is the amount of virtual pheromone on edge (i, j).

Two types of information are available to ants during their graph traversal, local and global, controlled by the parameters β and α respectively. Local information is obtained through a problem-specific heuristic measure. For the purposes of this paper, the fuzzy-rough dependency measure defined in equation (6) is used for this. The extent to which the measure influences an ant's decision to traverse an edge is controlled by the parameter β. This will guide ants towards paths that are likely to result in good solutions. Global knowledge is also available to ants through the deposition of artificial pheromone on the graph edges by their predecessors over time. The impact of this knowledge on an ant's traversal decision is determined by the parameter α. Good paths discovered by past ants will have a higher amount of associated pheromone. How much pheromone is deposited, and when, is dependent on the characteristics of the problem. No other local or global knowledge is available to the ants in the standard ACO model, though the inclusion of such information by extending the ACO framework has been investigated [1]. The choice of α and β is determined experimentally.

Selection Process. The ACO feature selection process begins with the generation of a number of ants, k, which are then placed randomly on the graph (i.e. each ant starts with one random feature). Alternatively, the number of ants to place on the graph may be set equal to the number of features within the data; each ant starts path construction at a different feature. From these initial positions, they traverse edges probabilistically until a traversal stopping criterion is satisfied. The resulting subsets are gathered and then evaluated. If an

optimal subset has been found or the algorithm has executed a certain number of times, then the process halts and outputs the best feature subset encountered. If neither condition holds, then the pheromone is updated, a new set of ants are created and the process iterates once more.

Complexity Analysis. The time complexity of the ant-based approach to feature selection is $O(IAk)$, where I is the number of iterations, A the number of original features, and k the number of ants. In the worst case, each ant selects all the features. As the heuristic is evaluated after each feature is added to the reduct candidate, this will result in A evaluations per ant. After one iteration in this scenario, Ak evaluations will have been performed. After I iterations, the heuristic will be evaluated IAk times.

Pheromone Update. Depending on how optimality is defined for the particular application, the pheromone may be updated accordingly. To tailor this mechanism to find fuzzy-rough set reducts, it is necessary to use the fuzzy-rough dependency measure as the stopping criterion. This means that an ant will stop building its feature subset when the dependency of the subset reaches the maximum for the dataset. The pheromone on each edge is then updated according to the following formula:

$$\tau_{ij}(t+1) = (1-\rho).\tau_{ij}(t) + \Delta\tau_{ij}(t) \tag{8}$$

where

$$\Delta\tau_{ij}(t) = \sum_{k=1}^{n} (\gamma'(S^k)/|S^k|) \tag{9}$$

This is the case if the edge (i, j) has been traversed; $\Delta\tau_{ij}(t)$ is 0 otherwise. The value ρ is a decay constant used to simulate the evaporation of the pheromone, S^k is the feature subset found by ant k. The pheromone is updated according to both the fuzzy-rough measure of the "goodness" of the ant's feature subset (γ') and the size of the subset itself. By this definition, all ants update the pheromone. Alternative strategies may be used for this, such as allowing only the ants with the currently best feature subsets to proportionally increase the pheromone.

To show the utility of fuzzy-rough feature selection and to compare the hill-climbing and ant-based fuzzy-rough approaches, the two methods are applied as pre-processors within a webpage classification system. Both methods preserve the semantics of the surviving features after removing redundant ones. This is essential in satisfying the requirement of user readability of the generated knowledge model, as well as ensuring the understandability of the pattern classification process.

4 Web Classification

There are an estimated 1 billion webpages available on the WWW with around 1.5 million webpages being added every day. The task to find a particular webpage, which satisfies a user's requirements by traversing hyper-links, is very

difficult. To aid this process, many web directories have been developed - some rely on manual categorization whilst others make decisions automatically. However, as webpage content is vast and dynamic, manual categorization is becoming increasingly impractical. Automatic web content categorization is therefore required to deal with these problems.

Information can be structured within a webpage that may indicate a relatively higher or lower importance of the contained text. For example, terms appearing within a <TITLE> tag would be expected to be more informative than the majority of those appearing within the document body at large. Because of this, keywords are weighted not only according to their statistical occurrence but also to their location within the document itself. These weights are almost always real-valued, which can be a problem for most feature selectors unless data discretization takes place (a source of information loss). This motivates the application of FRFS techniques to this domain.

Initial investigations have been carried out in this area [7], however these employed simplistic methods for classification - the vector space model and the boolean inexact model. The work presented here investigates the utility of more powerful approaches for this task, with the novel use of ACO-assisted feature selection.

4.1 System Overview

A key issue in the design of the system was that of modularity; it should be able to integrate with existing (or new) techniques. The current implementations allow this flexibility by dividing the overall process into several independent sub-modules:

- *Keyword Acquisition.* From the collected webpages, keywords/terms are extracted and weighted according to their perceived importance, resulting in a new dataset of weight-term pairs. These weights are almost always real-valued, hence the problem serves well to test the present work. For this, the TF-IDF metric [12] is used.
- *Keyword Selection.* As the newly generated datasets are too large, mainly due to keyword redundancy, a dimensionality reduction step is carried out using the techniques described previously.
- *Keyword Filtering.* Employed only in testing, this simple module filters the keywords obtained during acquisition, using the reduct generated in the keyword selection module.
- *Classification.* This final module uses the reduced dataset to perform the actual categorization of the test data. Four classifiers were used for comparison, namely C4.5 [10], JRip [3], PART [13] and a fuzzy rule inducer, QSBA [11]. Both JRip and PART are available from [14].

  C4.5 creates decision trees by choosing the most informative features and recursively partitioning the data into subtables based on their values. Each node in the tree represents a feature with branches from a node representing the alternative values this feature can take according to the current

subtable. Partitioning stops when all data items in the subtable have the same classification. A leaf node is then created, and this classification assigned.

JRip learns propositional rules by repeatedly growing rules and pruning them. During the growth phase, antecedents are added greedily until a termination condition is satisfied. Antecedents are then pruned in the next phase subject to a pruning metric. Once the ruleset is generated, a further optimization is performed where rules are evaluated and deleted based on their performance on randomized data.

PART generates rules by means of repeatedly creating partial decision trees from data. The algorithm adopts a separate-and-conquer strategy in that it removes instances covered by the current ruleset during processing. Essentially, a rule is created by building a pruned tree for the current set of instances; the leaf with the highest coverage is made into a rule.

QSBA induces fuzzy rules by calculating the fuzzy subsethood of linguistic terms and the corresponding decision variables. These values are also weighted by the use of fuzzy quantifiers. This method utilises the same fuzzy sets as those involved in the fuzzy-rough reduction methods.

4.2 Experimentation and Results

Initially, datasets were generated from large textual corpora collected from Yahoo [15] and separated randomly into training and testing sets, maintaining class distributions. Each dataset is a collection of web documents. Five classification categories were used, namely Art & Humanity, Entertainment, Computers & Internet, Health, Business & Economy. A total of 280 web sites were collected from Yahoo categories and classified into these categories. From this collection of data, the keywords, weights and corresponding classifications were collated into a single dataset.

Table 1 shows the resulting degree of dimensionality reduction, performed via selecting informative keywords, by the standard fuzzy-rough method (FRFS) and the ACO-based approach (AntFRFS). AntFRFS is run several times, and the results averaged both for classification accuracy and number of features selected. It can be seen that both methods drastically reduce the number of original features. AntFRFS performs the highest degree of reduction, with an average of 14.1 features occurring in the reducts it locates.

Table 1. Extent of feature reduction

Original	FRFS	AntFRFS
2557	17	14.10

To see the effect of dimensionality reduction on classification accuracy, the system was tested on the original training data and a test dataset. The results are summarised in table 2. Clearly, the fuzzy-rough methods exhibit better

resultant accuracies for the test data than the unreduced method for all classifiers. This demonstrates that feature selection using either FRFS or AntFRFS can greatly aid classification tasks. It is of additional benefit to rule inducers as the induction time is decreased and the generated rules involve significantly fewer features. AntFRFS improves on FRFS in terms of the size of subsets found and resulting testing accuracy for QSBA and PART, but not for C4.5 and JRip. The challenging nature of this particular task can be seen in the overall low accuracies produced by the classifiers (perhaps due to overfitting), though improved somewhat after feature selection. Both fuzzy-rough approaches require a reasonable fuzzification of the input data, whilst the fuzzy sets are herein generated by simple statistical analysis of the dataset with no attempt made at optimizing these sets. A fine-tuned fuzzification will certainly improve the performance of FRFS-based systems. Finally, it is worth noting that the classifications were checked automatically. Many webpages can be classified to more than one category, however only the designated category is considered to be correct here.

Table 2. Classification performance

	Original		FRFS		AntFRFS	
Classifier	Train	Test	Train	Test	Train	Test
C4.5	95.89	44.74	86.30	57.89	81.27	48.39
QSBA	100.0	39.47	82.19	46.05	69.86	50.44
JRip	72.60	56.58	78.08	60.53	64.84	51.75
PART	95.89	42.11	86.30	48.68	82.65	48.83

5 Conclusion

This paper has presented an ACO-based method for feature selection, with particular emphasis on fuzzy-rough feature selection. This novel approach has been applied to aid classification of web content, with very promising results. In all experimental studies there has been no attempt to optimize the fuzzifications or the classifiers employed. It can be expected that the results obtained with such optimization would be even better than those already observed.

There are many issues to be explored in the area of ACO-based feature selection. The impact of parameter settings should be investigated - how the values of α, β and others influence the search process. Other important factors to be considered include how the pheromone is updated and how it decays. There is also the possibility of using different static heuristic measures to determine the desirability of edges. A further extension would be the use of dynamic heuristic measures which would change over the course of feature selection to provide more search information. Future work will inlcude experimental investigations comparing current rough set-based methods (such as [6,16]) with the proposed approach on benchmark data.

References

1. E. Bonabeau, M. Dorigo, and G. Theraulez. Swarm Intelligence: From Natural to Artificial Systems. Oxford University Press Inc., New York, NY, USA. 1999.
2. A. Chouchoulas and Q. Shen. Rough set-aided keyword reduction for text categorisation. Applied Artificial Intelligence, Vol. 15, No. 9, pp. 843–873. 2001.
3. W.W. Cohen. Fast effective rule induction. In Machine Learning: Proceedings of the 12th International Conference, pp. 115–123. 1995.
4. M. Dash and H. Liu. Feature Selection for Classification. Intelligent Data Analysis, Vol. 1, No. 3, pp. 131–156. 1997.
5. D. Dubois and H. Prade. Putting rough sets and fuzzy sets together. In R. Slowinski (Ed.), Intelligent Decision Support, Kluwer Academic Publishers, pp. 203–232. 1992.
6. J. Han, X. Hu, and T.Y. Lin. Feature Subset Selection Based on Relative Dependency between Attributes. Rough Sets and Current Trends in Computing: 4th International Conference (RSCTC 2004), pp. 176–185. 2004.
7. R. Jensen and Q. Shen. Fuzzy-rough attribute reduction with application to web categorization. Fuzzy Sets and Systems, Vol. 141, No. 3, pp. 469–485. 2004.
8. R. Jensen and Q. Shen. Semantics-Preserving Dimensionality Reduction: Rough and Fuzzy-Rough Based Approaches. IEEE Transactions on Knowledge and Data Engineering, Vol. 16, No. 12, pp. 1457–1471. 2004.
9. Z. Pawlak. Rough Sets: Theoretical Aspects of Reasoning About Data. Kluwer Academic Publishing, Dordrecht. 1991.
10. J.R. Quinlan. *C4.5: Programs for Machine Learning*. The Morgan Kaufmann Series in Machine Learning. Morgan Kaufmann Publishers, San Mateo, CA. 1993.
11. K. Rasmani and Q. Shen. Modifying weighted fuzzy subsethood-based rule models with fuzzy quantifiers. In Proceedings of the 13th International Conference on Fuzzy Systems, pp. 1687–1694. 2004.
12. G. Salton and C. Buckley. Term Weighting Approaches in Automatic Text Retrieval. Information Processing and Management, Vol. 24, No. 5, pp. 513–523. 1988.
13. I.H. Witten and E. Frank. Generating Accurate Rule Sets Without Global Optimization. In Machine Learning: Proceedings of the 15th International Conference, Morgan Kaufmann Publishers, San Francisco. 1998.
14. I.H. Witten and E. Frank. Data Mining: Practical machine learning tools with Java implementations. Morgan Kaufmann Publishers, San Francisco. 2000.
15. Yahoo. `www.yahoo.com`
16. J. Yao and M. Zhang. Feature Selection with Adjustable Criteria. Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, 10th International Conference (RSFDGrC 2005), pp.204–213. 2005.
17. L.A. Zadeh. Fuzzy sets. Information and Control, 8, pp. 338–353. 1965.